A Graph-Theoretic Method for Analysing Structural Protein Variability in Viral Variants

L. Praveenkumar[1], G. Mahadevan[2], A. A. Navish[3*]

[1,2,3*] Department of Mathematics, The Gandhigram Rural Institute (Deemed to be University), Dindigul, TamilNadu, India - 624 302.

*** Corresponding Author**
Email: aa.navish2@gmail.com

Abstract: This study presents a graph-theoretic framework to analyze codon networks derived from SARS-CoV-2 spike protein sequences. By applying Minimum Connected Dominating Sets (MCDS) and community detection methods, we identify key codons that maintain both global connectivity and local structural integrity. Centrality measures are used to determine the most influential codons, highlighting their importance in network stability. Additionally, statistical analysis provides insights into the structural robustness of the spike protein across different variants.

Keywords: Codon network, Domination, MCDS, Centrality, Community, Protein targets

1 Introduction

The emergence and rapid evolution of SARS-CoV-2 highlight the urgent need for advanced analytical strategies to study its spike (S) protein. This protein binds to the human ACE2 receptor and plays a central role in viral entry, making it a primary target for vaccines and therapeutics [3]. Variants such as Alpha, Beta, Gamma, and Delta exhibit critical mutations in the spike protein that affect transmissibility, challenge vaccine efficacy, and contribute to immune escape [1], [2], [8].

The spike (S) protein of SARS-CoV-2 (ref figure 1) helps the virus enter human cells by binding to the ACE2 receptor, making it a key target for vaccines and treatments (Polack et al., [9]; Baden et al., [2]). mRNA vaccines like Pfizer-BioNTech and Moderna show around 95% effectiveness, while AstraZeneca and Johnson & Johnson use viral vectors and have shown 76% and 66% effectiveness, respectively (Voysey et al., [11] 2021; Sadoff et al., [10] 2021). Treatments like monoclonal antibodies such as casirivimab/imdevimab and bamlanivimab/etesevimab target the spike protein and help reduce the virus in the body. New variants, such as Alpha, Beta and Delta have spike protein mutations that make the virus spread faster and sometimes avoid immune defenses [1], [2], [8], [12]. As a result, researchers are working on small molecules to block the spike protein and stop the virus from infecting cells.

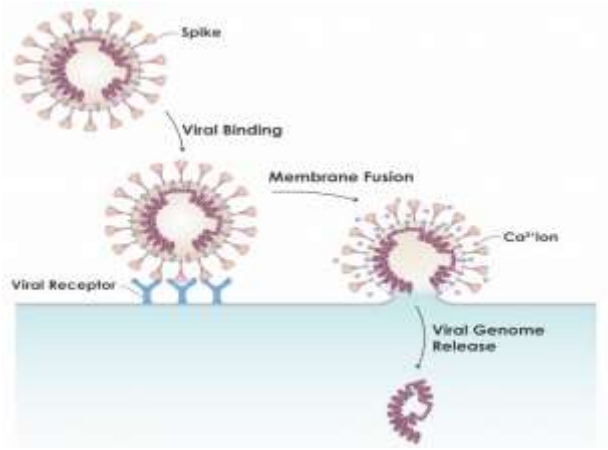


Figure 1: SARS Cov-2 entry in host cell

In this row, graph-theoretical methods have proven effective in studying protein structures and functions. Previous studies identified key residues and interactions [4], [7], explored spike variants using graph embeddings [5] and applied domination-based strategies in network modeling [6]. Consequently, in this work we apply a graph-theoretic framework to codon networks from SARS-CoV-2 spike sequences. By integrating MCDS, community detection, and centrality measures, we identify codons essential for network connectivity and modular structure. Our analysis highlights structurally and functionally important codons and offers insights into spike protein robustness across variants, aiding targeted antiviral development. To provide a brief overview of our work, the flowchart is presented. To provide a brief overview of our work, the flowchart illustration is presented in figure 2.

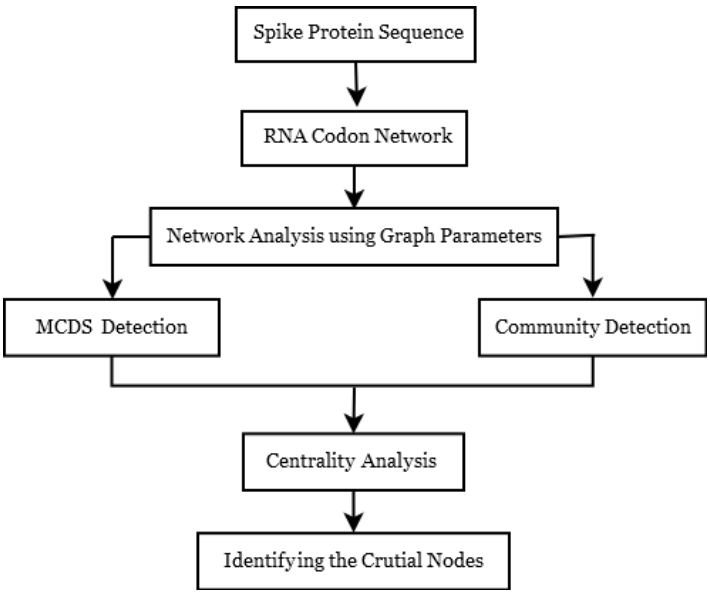


Figure 2: The flowchart illustration of our work

2 Methodologies

2.1 Decoding Protein Sequences into Codons

Amino acids are the basic units of proteins, combining in specific sequences to perform vital biological functions such as catalysis, transport, and structural support. Translating protein sequences into codons, where each amino acid corresponds to a three-nucleotide sequence, is essential for understanding gene expression and protein synthesis. This process allows researchers to decode the genetic basis of protein function and design targeted treatments like monoclonal antibodies and therapeutic enzymes. To demonstrate this, a simple protein sequence is converted into its corresponding codons as follows.

MYSFVSEETGTLIVNSVLLFLAFVVFLLVTLAILTALRLCAYCCNIVNVSLVKPSFYVYSRVKNLNSS RVPDLLV

AUGUACUUCU AUG UAC UUC UGU UUC GUA UCU UAA GAA UAC UGU UAC UUA UAU CUC UUU UUA UUA GCU UAC UAU GUA GCU UAC UUU CCA UAC AUC UUU CAG CCA CAA GCA UAA GAC GUA UAA UAC CCA UAA CCA UAA UUA UAA

2.2 Construction of Codon Graphs

A codon graph visually represents the interactions between codons within a protein sequence, simplifying complex biological data for analysis. It enables the identification of critical regions essential to protein function and provides insight into evolutionary differences among viral variants.

Definition 2.2.1 *Mathematically a* ***codon graph*** $G = (V, E)$ *is a mathematical representation of codon sequences from a protein sequence, where*

- *V is the set of unique RNA codons (triplets of nucleotides) derived from the sequence:*
  $$V = \{c_1, c_2, \dots, c_n\}.$$
  Here c_i *is a codon derived from the amino acids of the protein sequence.*
- *E is the set of undirected edges representing co-occurrence of codons within the same sequence context, regardless of order:*
  $E = \{\{u, v\} | u, v \in V \text{ and } u \text{ and } v \text{ are both present in the RNA sequence}\}$

2.3 Implementing Graph Parameters in Codon Graphs

In our study, the codon graph has the same set of codons (nodes) across variants, but the connections (edges) differ based on codon interactions. This makes each variant's graph structure unique.

Minimum Connected Dominating Set

Among various domination parameters, the MCDS is favored for its balance between minimal size and high connectivity. It identifies a compact, connected group of key codons that form the backbone of the codon network. These codons preserve essential interactions

while simplifying network analysis. Because modifications to these codons can significantly impact protein structure and function, they serve as valuable targets for drug development and genetic engineering.

Definition 2.3.1 *A set $D \subseteq V$ is a **dominating set** if every vertex $v \in V$ is either in D or adjacent to at least one vertex in D. In other words, for every vertex $v \notin D$, there exists a vertex $u \in D$ such that $(u, v) \in E$. The set D is a **connected dominating set** if the subgraph induced by D is connected.*

*A connected dominating set D is a **minimum connected dominating set** if it has the smallest possible cardinality among all connected dominating sets in the graph. This means that there is no other connected dominating set in H with fewer vertices than D.*

Centralities

Centralities in a codon graph help identify key codons important for protein structure and function. Highly central codons are well-connected and often crucial for protein stability, synthesis, and folding. These codons tend to be conserved across species, highlighting their evolutionary significance.

Definition 2.3.2 *If A is the adjacency matrix of a graph, the **eigenvector centrality** ($\mathcal{EC}$) x_i of a node i is given by the solution to the following eigenvector equation:*

$$A\boldsymbol{x} = \lambda\boldsymbol{x}$$

where $\boldsymbol{x}$ is the eigenvector associated with the largest eigenvalue λ of the matrix A. The centrality score for node i is the i-th component of the eigenvector $\boldsymbol{x}$.

Definition 2.3.3 *The **degree centrality** ($\mathcal{DC}$) of a node v_i in a graph is defined as:*

$$\mathcal{DC}(v_i) = \deg(v_i) = |\{v_j \in V : (v_i, v_j) \in E\}|$$

where $deg(v_i)$ represents the degree of node v_i (the number of edges connected to v_i). This measure indicates the immediate connections a node has, with higher values suggesting greater influence or importance within the network.

Definition 2.3.4 *The **betweenness centrality** ($\mathcal{BC}$) of a node v_i in a graph is defined as*

$$\mathcal{BC}(v_i) = \sum_{s \neq v_i \neq t} \frac{\sigma_{sp}(v_i)}{\sigma_{sp}}$$

where σ_{sp} is the number of shortest paths between nodes s and t and $\sigma_{sp}(v_i)$ is the number of those shortest paths that pass through node v_i.

Definition 2.3.5 *The **closeness centrality** ($\mathcal{CC}$) of a node v_i in a graph is defined as:*

$$\mathcal{CC}(v_i) = \frac{1}{\sum_{j \neq v_i} d(v_i, j)}$$

where $d(v_i, j)$ is the shortest path distance between nodes v_i and j. This measure indicates how quickly a node can reach all other nodes in the network, with higher values representing nodes that are more centrally located and can access other nodes more efficiently.

Communities

Analyzing the community structures of SARS CoV-2 and its variants, beyond just MCDS nodes offers a deeper view into network organization. While MCDS nodes highlight essential elements, a full community analysis reveals the supportive and bridging codons

that enhance each variant's resilience, adaptability and connectivity. This approach uncovers structural patterns that influence mutation behavior and transmissibility. Here, the InfoMap algorithm is applied. This algorithm detects communities through random walk modeling by minimizing the path description lengths. It effectively partitions nodes into densely connected clusters, revealing shared characteristics within complex networks.

Definition 2.3.6 *The main objective of the InfoMap algorithm is to minimize the expected description length L of the random walk paths and is expressed as*

$$L = -\sum_{k}\left(\frac{N_k}{N} log \frac{N_k}{N} + \sum_{i\in C_k}\sum_{j\in C_k} P(i\to j) log P(i\to j)\right)$$

where, N_k is the number of visits to community C_k and N is the total number of visits across all communities and P represents the transition probability. A community $C \subseteq V$ is defined as a subset of nodes exhibiting a higher density of connections among themselves than with nodes outside the community. The goal is to partition the graph G into disjoint communities $C_1, C_2, \dots, C_m$ that minimize the description length L.

2.4 Statistical Measure

In this study, we focus on the numeric values obtained from the Jaccard similarity calculations to explore the variations among the MCDS and community nodes of the spike proteins across SARS CoV-2 variants. By analyzing these values, we aim to highlight the degree of similarity or dissimilarity between the variants. This analysis can offer insights into their evolutionary relationships and functional adaptations. The Jaccard similarity coefficient is defined mathematically as follows.

Definition 2.4.1 *Given two sets A and B, the **Jaccard similarity coefficient** $J(A,B)$ is expressed as:*

$$J(A,B) = \frac{|A\cap B|}{|A\cup B|}$$

where, $|A \cap B|$ is the cardinality (size) of the intersection of sets A and B, representing the number of elements that are present in both sets; $|A \cup B|$ is the cardinality of the union of sets A and B, representing the total number of distinct elements that are present in either set.

A value of 0 indicates that the two sets do not share any common elements (i.e., they are completely disjoint). A value of 1 indicates that the two sets are identical (i.e., they contain exactly the same elements).

3 Results and Discussion

The protein sequences utilized in this study are obtained from the RCSB Protein Data Bank (https://www.rcsb.org/). Then the sequences are transformed into a codon network, and all graph-related analyses are conducted using Python. The examined spike protein structures and their corresponding PDB IDs are: Original (7DDN), Alpha (7R13), Beta (7VX1), Gamma (8DLQ) and Delta (8HRJ).

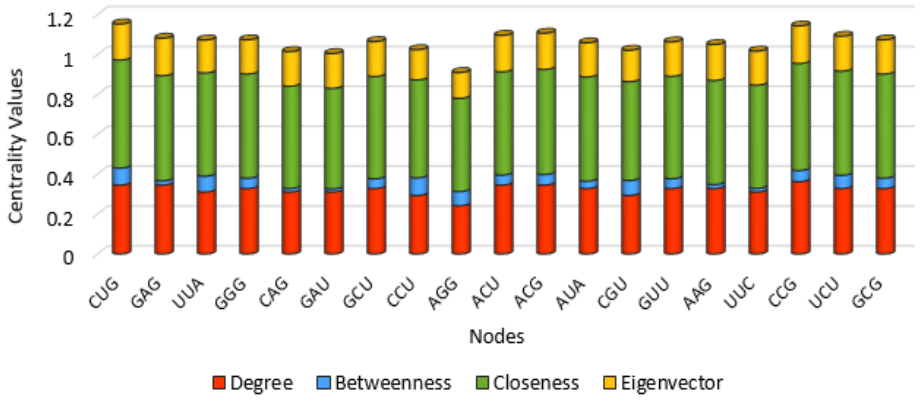
The obtained protein sequence is first translated into a codon sequence based on the genetic code, where each amino acid in the protein is encoded by one or more triplet codons and the codon network is constructed.

From our analysis of the codon network among distinct SARS CoV-2 variants, we found that while the number of nodes in the codon graphs remains consistent, the number of edges varies. This indicates that, although the sequences encode the same amino acids, they may employ different codon choices. These variations reveal diverse patterns of redundancy and connectivity resulting from mutations.

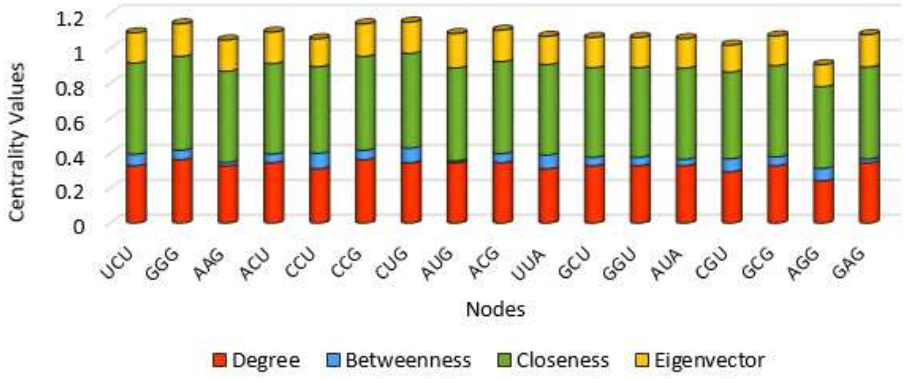
To further investigate the codon network of SARS CoV-2 variants, we employed both local and global graph-theoretic approaches. This dual strategy allows us to understand not only the specific effects of mutations on codon usage within individual variants but also the broader structural implications for the entire network.

In the first step, the network nodes with maximum centrality are identified and their bar chart representations based obtaines measures are shown in the figure 3.

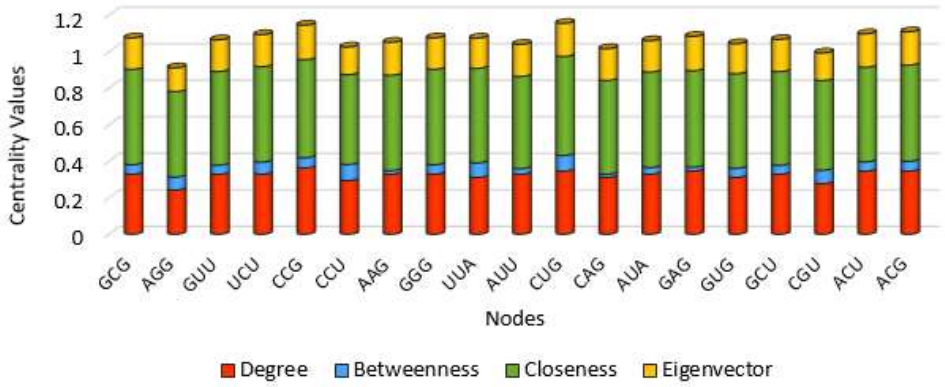
SARS Cov-2



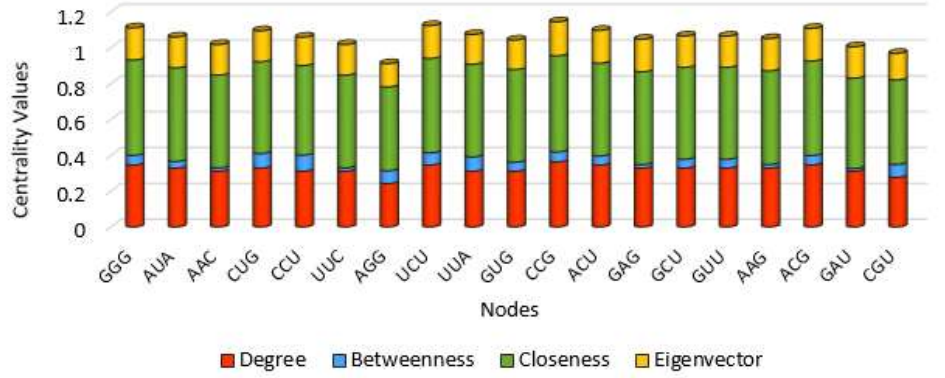
Alpha



Beta



Gamma



Delta

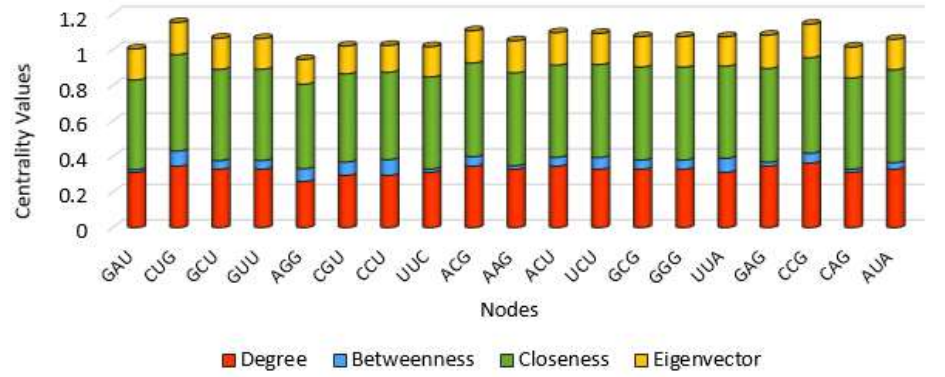


Figure 3: Bar chart interpretation of the centralities of the top nodes belonging to various spike codon networks

Next, we move on to the analysis of the MCDS and community structure. In this step, we assess the influence of the identified central nodes on both network connectivity and the formation of cohesive communities.

Variants	Nodes in MCDS	MCDS nodes with Max centralities
SARS CoV-2	UUA, GUC, GCU, UUG, AGG, GGU, CUC, CCU, CUU, CGG, AUU, GCC, ACC, UCC, ACU, UCU, CGA, GGC,	UCU, AGG, GCU, GUU, CCU, UUA, CCG, ACU

	CCG, AGA, GUU	
Alpha	CCG, ACU, CGG, AGA, GCC, UUG, AGG, GCU, ACC, CUU, UCU, UUA, CGA, GUU, GUC, UCC, GGC, GGU, CUC, AUU, CCU	AGG, UCU, GCU, CCG, UUA, ACU, CCU, GUU
Beta	GGU, CUC, CCU, UCU, GCC, UUG, AGG, ACU, GUU, CCG, AGA, AUU, GGC, CGA, CGG, CUU, GUC, UUA, ACC, UCC, GCU	CCU, GUU, AUU, UUA, ACU, UCU, AGG, CCG, GCU
Gamma	CGG, CUU, GGU, GGC, CUC, UUA, ACC, ACU, AUU, UCU, CGA, CCG, GUU, CCU, UCC, AGG, UUG, GCC, AGA, GCU, GUC	GCU, CCU, UCU, GUU, CCG, UUA, ACU, AGG
Delta	GUC, ACU, CGG, CCG, CGA, CUA, GUU, AGG, UCC, UCU, ACC, CUG, GCU, GGU, AUA, CUU, UCA, AGA, CCA, GGG, CUC, GCC	GGU, AGG, CUG, GCU, GGG, UCU, AUA, ACU, CCG

Table 1: SARS CoV-2 Variants and their corresponding MCDS

The table 1 presents the common nodes identified across all the MCDS of SARS CoV-2 variants, including **CUC, GCC, UCU, CUU, AGA, AGG, CCG, GCU, CGA, CGG, ACU, ACC, UCC** represent crucial elements that are integral to the virus's biological functions. Targeting these conserved elements can inform vaccine design and lead to broad-spectrum antiviral treatments effective against multiple strains. Additionally, these nodes can serve as biomarkers for monitoring infections and guiding combination therapies, enhancing treatment efficacy and reducing resistance development.

Instead of examining the entire network, concentrating on community studies will reduce complexity and facilitate the identification of essential nodes and interactions. It will be more suitable for a wide and complex network. Since the MCDS for the Alpha, Beta and Gamma variants is identical to that of SARS CoV-2, we shifted our focus to the community analysis of these variants to gain deeper insights. Surprisingly, we discovered that the connected communities of the variants are largely identical, with the exception of one huge community. Therefore, analyzing this larger community should provide adequate information for a more comprehensive understanding of the variants.

Variants	**Community nodes with Max centrality**	**Community nodes involved in MCDS with Max centrality**
SARS CoV-2	CAG, CCU, CGU, ACU, UCU, UUC, UUA, ACG, AGG, CUG, AAG, GGG, GAG, GAU, GCG, CCG, AUA, GCU, GUU	AGG, ACU, CCU, GUU, GCU, UUA, CCG, UCU
Alpha	ACU, UUA, CGU, GGG, GUU, CCG, UCU, AAG, CAG, CCU, GAU, CUG, GCG, AUA, GAG, GCU, AGG, ACG, UUC	AGG, ACU, CCU, GUU, GCU, UUA, CCG, UCU

Beta	AUA, GUG, GCG, UUA, CCU, CGU, AGG, ACU, AUU, GUU, CAG, UCU, CUG, AAG, CCG, GGG, GCU, GAG, ACG	AGG, ACU, CCU, GUU, GCU, AUU, UUA, CCG, UCU
Gamma	CCG, UCU, GCU, CGU, UUA, AAG, AUA, CCU, AAC, GAG, CUG, GUU, GUG, ACU, GAU, AGG, ACG, GGG, UUC	AGG, ACU, CCU, GUU, GCU, UUA, CCG, UCU
Delta	CUG, GCU, ACG, CCU, AUA, CCG, GAG, UCU, GCG, GGU, GGG, ACU, UUA, AAG, AUG, CGU, AGG	AGG, GGU, ACU, CUG, GGG, GCU, AUA, CCG, UCU

Table 2: Large Community analysis

Table 2 shows the link between MCDS and community nodes, with 50% of MCDS belonging to the large community. It also lists community nodes with maximum centrality, separating those that are part of the MCDS in the last column is identical to Table 1's last column. This overlap indicates that the most central nodes are also core members of the largest community, emphasizing their importance in maintaining network structure and connectivity. Targeting them could enhance both network efficiency and stability.

From Tables 1 and 2, nodes like **AGG, ACU, CCU, GUU, GCU, UUA, CCG, UCU** consistently appear across all networks, communities, and MCDS with high centrality. Their broad presence underlines their key role in viral codon structure and function. Thus, targeting these nodes may support the development of broad-spectrum antiviral therapies.

Variants	**Unique MCDS Nodes of Variants with SARS CoV-2 MCDS**	**Unique Community Nodes of Variants with SARS CoV-2 Community**
Alpha	Nil	ACA, ACC
Beta	Nil	GCC, GCA
Gamma	Nil	GGC, GGA
Delta	AUA, CUG, CCA, GGG, CUA, UCA	GUC, ACA, ACC, GUA

Table 3: Unique MCDS and Community Nodes of Variants Compared to the Original Strain

Table 3 shows that Alpha, Beta, and Gamma have the same MCDS nodes as the original SARS-CoV-2, meaning there are no major changes. However, each of them has two new community nodes, showing small differences. Delta is more different, with six new MCDS nodes and four new community nodes. These new nodes in Delta are all part of its large communities. This makes those communities important targets.

Finally, a numeric analysis were given based on SARS CoV-2 vs. variants. To verify whether each variant is different from the other, the Jaccard index is obtained and presented in the tables 4 and 5. These two tables summarize the Jaccard similarity coefficients among

SARS CoV-2 and its variants (Alpha, Beta, Gamma and Delta) based on MCDS and community nodes of codon networks.

Variants	SARS CoV-2	Alpha	Beta	Gamma	Delta
SARS CoV-2	1.00000	1.00000	1.00000	1.00000	0.59259
Alpha	1.00000	1.00000	1.00000	1.00000	0.59259
Beta	1.00000	1.00000	1.00000	1.00000	0.59259
Gamma	1.00000	1.00000	1.00000	1.00000	0.59259
Delta	0.59259	0.59259	0.59259	0.59259	1.00000

Table 4: Jaccard similarity for MCDS of considered various codon networks

Variants	SARS CoV-2	Alpha	Beta	Gamma	Delta
SARS CoV-2	1.00000	0.95000	0.95000	0.95000	0.90476
Alpha	0.95000	1.00000	0.90476	0.90476	0.95238
Beta	0.95000	0.90476	1.00000	0.90476	0.86364
Gamma	0.95000	0.90476	0.90476	1.00000	0.86364
Delta	0.90476	0.95238	0.86364	0.86364	1.00000

Table 5: Jacard similarity for communities of considered various codon networks

Table 5 shows that SARS CoV-2, Alpha, Beta and Gamma exhibit perfect similarity coefficients of 1.00000 with each other, indicating nearly identical genetic structures. In contrast, Delta displays lower similarity values (0.59259 and 0.61539, respectively), reflecting significant genetic divergence. Table 6 reinforces these findings, with high similarities (around 0.95) among the original strain and early variants, while Delta shows reduced similarities (0.90476 and 0.86364). From that, we can conclude that none of the communities are exactly the same.

4 Conclusion

In conclusion, our analysis of MCDS, community structures, and centrality in SARS-CoV-2 codon networks revealed key codons vital to viral function. Community analysis showed how mutations influence codon grouping and interaction, impacting variant behavior and treatment response. Identifying high-centrality codons in MCDS and major communities highlighted potential therapeutic targets. Targeting these codons may disrupt the spike protein network, aiding antiviral strategies. Jaccard similarity further revealed genetic relationships among variants, offering insight into their evolution and vulnerabilities. This framework supports the development of effective therapies and informed public health responses.

References

[1]. Ahmad, A., Fawaz, M. A. M., & Aisha, A. (2022). A comparative overview of SARS-Cov-2 and its variants of concern. Le Infezioni in Medicina, 30(3), 328.

[2]. Baden, L. R., El Sahly, H. M., Essink, B., Kotloff, K., Frey, S., Novak, R., ... & Zaks, T. (2021). Efficacy and safety of the mRNA-1273 SARS-Cov-2 vaccine. New England Journal of Medicine, 384(5), 403-416.

[3]. Dougan, M., Nirula, A., Azizad, M., Mocherla, B., Gottlieb, R. L., Chen, P., ... & Skovronsky, D. M. (2021). Bamlanivimab plus etesevimab in mild or moderate Covid-19. New England Journal of Medicine, 385(15), 1382-1392.

[4]. Ghosh, S., Kumar, G. V., Basu, A., & Banerjee, A. (2015). Graph theoretic network analysis reveals protein pathways underlying cell death following neurotropic viral infection. Scientific Reports, 5(1), 14438.

[5]. Guzzi, P. H., Lomoio, U., Puccio, B., & Veltri, P. (2022). Structural analysis of SARS-Cov-2 Spike protein variants through graph embedding. Network Modeling Analysis in Health Informatics and Bioinformatics, 12(1), 3.

[6]. Mahadevan, G., Ponnuchamy, T., Avadayappan, S., & Mishra, J. (2020). Application of eternal domination in epidemiology. In Mathematical modeling and soft computing in epidemiology (1st ed., pp. 25). CRC Press.

[7]. Manrique, P. D., Chakraborty, S., Henderson, R., Edwards, R. J., Mansbach, R., Nguyen, K., ... & Gnanakaran, S. (2023). Network analysis uncovers the communication structure of SARS-Cov-2 spike protein identifying sites for immunogen design. Iscience, 26(1).

[8]. Navish, A. A., & Uthayakumar, R. (2023). A comparative study on structural proteins of viruses that belong to the identical family. The European Physical Journal Special Topics, 232(7), 1051-1060.

[9]. Polack, F. P., Thomas, S. J., Kitchin, N., Absalon, J., Gurtman, A., Lockhart, S., ... & Gruber, W. C. (2020). Safety and efficacy of the BNT162b2 mRNA Covid-19 vaccine. New England Journal of Medicine, 383(27), 2603-2615.

[10]. Sadoff, J., Gray, G., Vandebosch, A., Cárdenas, V., Shukarev, G., Grinsztejn, B., ... & Douoguih, M. (2021). Safety and efficacy of single-dose Ad26.COV2.S vaccine against Covid-19. New England Journal of Medicine, 384(23), 2187-2201.

[11]. Voysey, M., Clemens, S. A. C., Madhi, S. A., Weckx, L. Y., Folegatti, P. M., Aley, P. K., ... & Bijker, E. (2021). Safety and efficacy of the ChAdOx1 nCoV-19 vaccine (AZD1222) against SARS-Cov-2: an interim analysis of four randomised controlled trials in Brazil, South Africa and the UK. The Lancet, 397(10269), 99-111.

[12]. Yi, H., Wang, J., Wang, J., Lu, Y., Zhang, Y., Peng, R., ... & Chen, Z. (2021). The emergence and spread of novel SARS-Cov-2 variants. Frontiers in Public Health, 9, 696664.